自然言語処理 70－10
（1989. 1. 20）

日本語「が」格の係り受け整合度の一計算法

田中 英輝　江原 暉将
NHK放送技術研究所

筆者らは、係り受け解析を行う際に必要となる文節間の整合度を設定する研究を行っている。これにはいくつかの方法が考えられるが、ここでは、内省による方法に着目し、「が」格で結合された主語名詞と自動詞の係り受け整合度を、9名の被験者による内省実験によって決定した。その結果、この方法の特長と問題点が明らかになった。また、この整合度を利用して主成分分析を行い、動詞を係り受けの観点で分類した。最後に、未知の単語の整合度を既知の整合度データから線形関数で予測する方法を示した。

A Method of Calculation for Matching Score between a Japanese Subject Noun and a Verb

Science and Technical Rsearch Laboratories of NHK
Hideki TANAKA and Terumasa EHARA

1-10-11 Kinuta Setagaya-ku Tokyo Japan

We have studied methods of fixing a matching score between two Japanese Bunsetu phrases for Dependency-relation-based parsing. In this paper, we calculated the matching scores between fifty verbs and sixteen nouns based on nine Japanese native speakers' introspective evaluation of eight hundred test sentences. Each sentence consisted of a subject noun ,a subject marker "Ga", and a verb. Checkers evaluated in four ranks the coexistence possibility of the noun and the verb in the test sentence. Principal component analysis was conducted on the matching scores to classify the verbs used in the experiments. We also predicted matching scores between unknown verbs and known nouns by a linear function on the basis of the learned matching scores .

1. はじめに

筆者らは現在、尾関によって考案された構文解析プログラム[1]で用いるために二文節間の係り受け整合度を設定する研究を行なっている。係り受け整合度を設定する方法として従来から3つの方法が提案されている。第1の方法は、既存のシソーラス[2][3]を利用するものであるが、それらには問題があることが指摘されており[4]、また係り受けという観点で分類されていないため我々の目的に即しているとは言い難い。

第2の方法はテキストデータから係り受け構造を抽出し整合度を決定する方法である[5][6]。この方法では、テキストデータ中には現れないけれども可能な係り受けが多く存在することが予想される。

第3の方法は内省による整合度の決定である[7][8]。ここでは、最後の方法に着目し、以下のような実験を行った。格助詞「が」を介して、主語名詞が動詞に係る係り受けに関し、内省による実験を行ない整合度を決定した。また共起という観点で語の分類を構成する際の知見を得るため、実験結果に対し主成分分析を行なった。さらに、未知の単語の整合度を既知データから予測することを試みた。

2. 内省実験

内省実験に用いたのは

「N」 + が + 「V」。

（名詞） （動詞終止形）

の形の文である。実験に使用した名詞は、広い意味範囲を代表するためIPAL[9]の意味素性分類から一つづつ単語を選択した。動詞はNHK編新用字用語辞典[10]の中から新明解国語辞典[11]に最重要と表示された動詞514個を選び、さらに類語新辞典の意味コードを与えた上で、自動詞を中心に、また多義性の少ない語を50個抽出した。使用名詞、使用動詞を表1、2に示す。

表1：名詞リスト

番号	単語	意味素性	番号	単語	意味素性
1	犬	ANI	11	ニュース	LIN
2	男性	HUM	12	美	CHA
3	企業	ORG	13	原因	REL
4	花	PLA	14	公園	LOC
5	頭	PAR	15	昨日	TIM
6	空	NAT	16	三日	QUA
7	紙	PRO			
8	風	PHE			
9	勉強	ACT			
10	心	MEN			

表2：動詞リスト

番号	単語	コード	番号	単語	コード
1	会う	781	26	壊れる	241
2	遊ぶ	891	27	しまう	282a
3	集まる	710	28	足りる	495
4	余る	261	29	近づく	215
5	現われる	233a	30	付く	224
6	生きる	071b	31	続く	225
7	急ぐ	296	32	飛ぶ	218
8	浮く	209	33	届く	223
9	怒る	472	34	止まる	200a
10	劣る	299a	35	直る	392
11	踊る	887	36	無くなる	260a
12	泳ぐ	898f	37	濁る	254b
13	折れる	244	38	残る	261
14	終わる	282a	39	始る	282
15	輝く	095c	40	はやる	761
16	隠れる	233b	41	光る	095c
17	欠ける	264b	42	広がる	243
18	固まる	251	43	太る	072b
19	偏る	296c	44	掘る	388a
20	変わる	277	45	曲がる	216a
21	切れる	261a	46	乱れる	278
22	崩れる	259	47	実る	057c
23	曇る	022c	48	行く	311a
24	苦しむ	492a	49	分かれる	220a
25	答える	429a	50	笑う	321

コードは類語新辞典の意味番号

このようにして構成した800個のテスト文を、日本語を母国語とする、27歳から46歳の9名の被験者（男性5名／大学卒、理科系：女性4名／大学卒、文科系）に提示して次の4段階で評価してもらった。

（1）○ ・・・ 自然に共起する。

（2）△ ・・・ 詩的、比喩的に共起する。

（3）× ・・・ 共起しない。

（4）？ ・・・ 不明。

評価の際に、動詞の意味を明らかにするため、被験者には類語新辞典の意味を記述した辞書を渡し、その中のどの意味で評価すべきかを示した。尚、？については、７２００文の内２９例と少なく、評価後のインタビュー調査の結果△と同じと見なせることが分かったため、△に繰り入れた。

3. 実験結果

各テスト文に対して９名の評価が得られた。例えば、

「犬　が　会う。」に対して｛△、×、△、○、○、△、△、△、△｝、

「男性　が　遊ぶ。」に対して｛○、○、○、○、○、○、○、○、○｝、

「空　が　集まる。」に対して｛×、×、×、×、×、×、×、×、×｝、

「犬　が　笑う。」に対して｛△、△、△、△、△、△、△、△、△｝

のような評価が得られた。このような評価を８００の文すべてについて、○の個数と×の個数で平面に記述した（図１）。

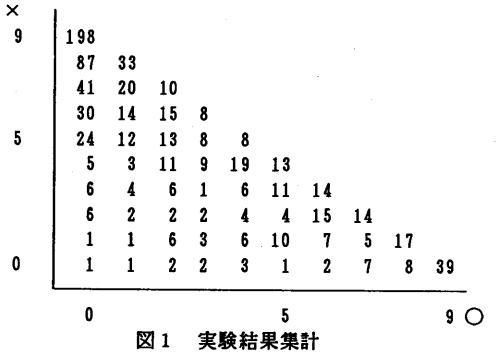

図１　実験結果集計

このグラフによると、全員で評価が一致したものは、○が３９、×が１９８、△が１となっている。すなわち全体の１／４程度は、全員の意見が一致したことがわかる。この中で×で一致したものが１９８と多い。これは特に、共起に自然性がない場合は意見が一致しやすいことを示している。

また「三日」、「昨日」といった名詞はどんな動詞ともほとんど共起しなかった。被験者によって評価が分かれた文についてその原因を調べるため、○と×の個数が１対８、２対７で評価が分かれた文（４３個）に対する追調査を行った。調査は、○と評価した人にその理由を尋ねることで行った。その結果以下のような８通りの理由に分類できることが分かった。

１）名詞への修飾語を付加する。
１４例
例、　花　が　止まる。
　　風にゆれていた花が止まった。

２）動詞への修飾語を付加する
１４例
例、　公園　が　集まる。
　　開発の結果、公園が集まる。

３）文脈の付加。
１８例
例、　昨日　が　曇る。
　　一昨日は夕焼けだったので、
　　昨日が曇るとは思わなかった。

４）助述成分の変更。
２３例
例、　公園　が　集まる。
　　このあたりには、公園が集まっている。
　　花　が　始まる。
　　森を出ると、花が始まった。

５）格助詞の変更。
３例
例、　公園　が　終わる。
　　その柵のところで、公園は終わっていた。

6）間接目的語の付加。
6例
例、 男性 が 広がる。
固まっていた男性が、庭に広がった。

7）省略があると考える。
7例
例、 頭 が 折れる。
頭の骨が折れる。

8）普通ではないが可能であると判断した文。 4例
例、 空 が 乱れる。

これらの例は、修飾語や間接目的語の付加によって共起の自然性が変化するため、二文節だけで係り受けの整合度を決定することに限界があることと、一方で、テキスト中に現れにくい係り受けを広い範囲の内省によって評価することが可能になっていることを示している。また助述成分や格助詞の変化と共起の自然性についてなんらかの規則があるかどうかの点は今後調査する必要がある。

4. 係り受けに基づく単語の分類

共起関係に基づいて単語を分類し、その特徴を明らかにできれば、未知の単語間の整合度を予測する場合に有効であると考えられる。そこで今回の実験結果を基に、名詞を軸として主成分分析を行い、動詞を分類し、その特徴を調査した。

まず各テスト文「N」が Vi」に対する9名の評価を、全員が○をつけると0点で、逆に全員が×をつけると10点になるように（1）式で表現した。この式の値を、整合度と呼ぶ。

$$m_{ij} = \frac{1}{m}(-5m_1+5m_2+5m) \quad \cdots (1)$$

m 被験者総数 9
m_1 ○の個数
m_2 ×の個数

こうすると各動詞に対して16次元の得点ベクトルが与えられるので、これを用いて主成分分析を行った。

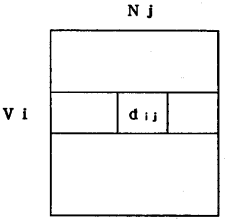

図2：動詞の得点ベクトル

軸の累積寄与率は、表3に示すとおりであり、3軸までで71%の情報が復元された。

表3：累積寄与率

次数	1	2	3	4	5
累積寄与率%	34	59	71	79	85

1－2軸散布図を図3に示す。意味的に近いと思われる「近づく」、「集まる」（類義語）：「始まる」、「終わる」（反義語）がそれぞれ近くに配置されている。このことは、共起関係に基づいた動詞の分類は、類義関係だけに着目するのでなく反義関係も考慮して行う必要があることを示している。

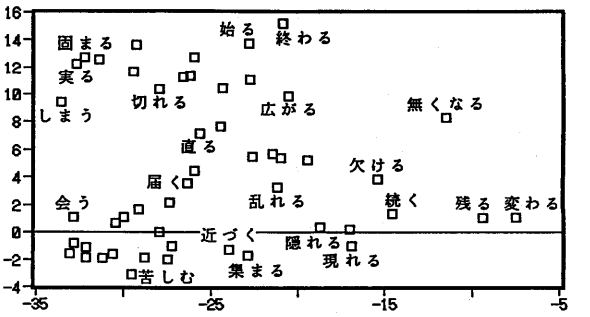

図3 動詞1－2軸散布図

分類の基準となった情報を詳細に調べるため、固有ベクトル中で、貢献の大きな名詞を調べ、軸の解釈を行った。固有ベクトルを図4に示す。第一固有ベクトルは、「犬」、「男性」の成分の絶対値が小さくその他の成分はおよそ同じ大きさである（「三日」、「昨日」を除く）。このことから第一軸は、非動物性主語一般との共起を示すと考えられる。第二固有ベクトルでは「犬」、「男性」の貢献が大きくなっており、第二軸は生物主語との共起を示すと解釈される。

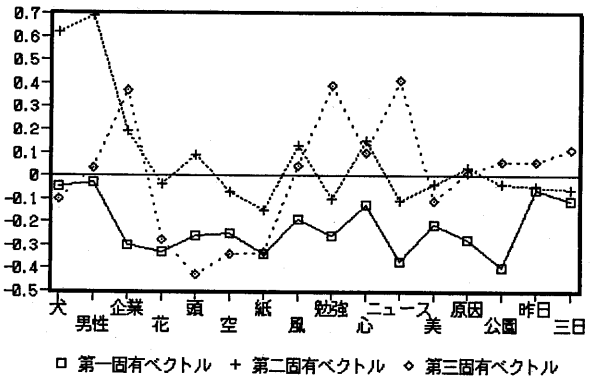

図4　固有ベクトル

5．整合度予測実験

1）問題の定式化

現在、図2の形で整合度マトリクス $M=(m_{ij})$ が与えられている。ここで新たに、動詞もしくは名詞が与えられた時、既知の整合度から新しい単語に対する整合度を予測することを考える。例えば、整合度マトリクスに対して動詞 V_{i+1} を一つ加えるとする。この時、全ての名詞との整合度を内省によって求めることは、労力がかかる。そこで、V_{i+1} と適当な基準名詞2つ（N_{n1}, N_{n2}）とで内省実験を行い、それぞれの整合度を求め、この整合度を利用して基準名詞以外の名詞との整合度を線形最小自乗予測する方法を検討した。具体的には以下の手続きで、新たな動詞 V_{i+1} と名詞 N_k との整合度を求めた。

ステップ1：学習データを使って、式（3）に従い線形予測係数 α_k、β_k、γ_k を求める。

ステップ2：動詞 V_{i+1} と基準名詞 N_{n1} と N_{n2} との整合度 $m_{i+1,n1}$、$m_{i+1,n2}$ を内省実験によって求める。

ステップ3：式（2）に従って動詞 V_{i+1} と名詞 N_k との整合度 $\hat{m}_{i+1,k}$ を計算する。（図5参照）

$$\hat{m}_{i+1,k} = \alpha_k * m_{i+1,n1} + \beta_k * m_{i+1,n2} + \gamma_k \quad \cdots (2)$$

N_{n1}：基準名詞1　N_{n2}：基準名詞2

$\hat{m}_{i+1,k}$ ：動詞 V_{i+1} と名詞 N_k（$k \neq n1, n2$）の予測整合度

α_k、β_k、γ_k：名詞 N_k（$k \neq n1, n2$）の線形予測係数

$$\sum_{i=1}^{L} (\hat{m}_{i,k} - m_{i,k})^2 \xrightarrow{\alpha_k, \beta_k, \gamma_k} \min \quad \cdots (3)$$

L：学習に用いる動詞の数

$$\sum_{i=1}^{5016} \sum_{k=1} (\hat{m}_{i,k} - m_{i,k})^2 \quad \cdots (4)$$

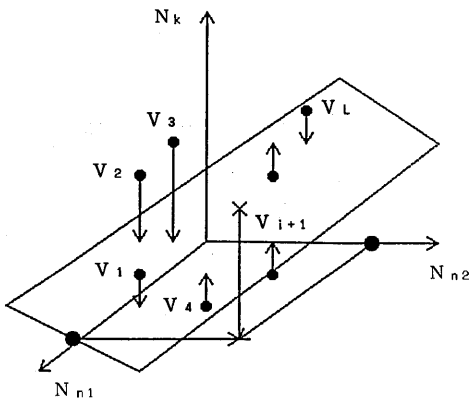

図5　整合度予測法

２）実験

動詞全てを学習データとした場合。基準名詞をすべてのやりかたで選択し（１２０通り）（４）式の予測自乗誤差合計が最小になる組合せを調べた。その結果「男性」と「公園」が最も良い結果となった（表４参照）。この場合、各m_{ij}の平均誤差は約１．９である。この組合せは動物性名詞と非動物性名詞になっている。

表４：５０動詞すべてで学習した場合の自乗誤差合計

名詞の組合せ	予測自乗誤差合計
男性、公園	２９６４．６
紙、　勉強	３８８８．２
犬、　男性	４２５３．０
昨日、三日	４９５１．６

３）学習数と誤差について。

予測誤差合計と学習データ数の関係を調べる実験を行った。学習する動詞の数を整合度テーブルの先頭から順に増やし、それぞれの場合で学習データ、未知データ両方の整合度を予測した。誤差は、（４）式、すなわち学習データ、未知データの合計で評価した。基準名詞としては「男性」と「公園」、「紙」と「勉強」、「犬」と「男性」、「昨日」と「三日」を採用した。結果をグラフにしたのが図６である。この図から明らかなように、基準名詞の種類にかかわらず、１５個程度学習すると誤差の降下は止まっている。

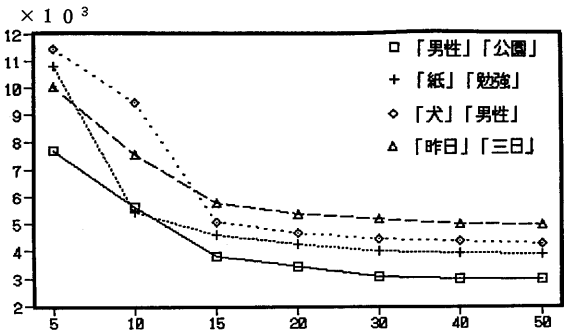

図６　動詞学習数対自乗誤差曲線

次に、予測誤差の度数分布を図７に示す。この図は、２０の動詞を学習した場合であり、全データ、学習データ、未学習データのそれぞれの領域で、実測整合度と予測整合度の差を計算したものを示している。

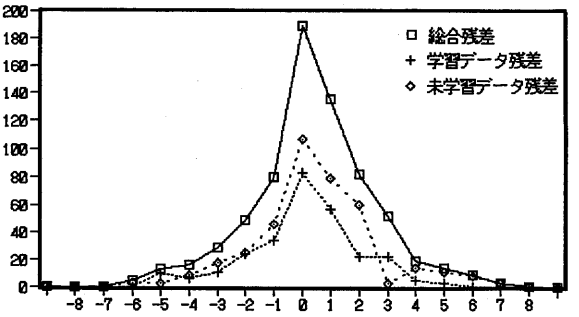

図７　予測誤差度数分布

最後に、学習数の増加の効果を個別の動詞について観察するため「会う」、「曇る」、「残る」という動詞に着目し、各名詞との予測整合度および実測整合度の変化をグラフにした。図８、図９、図１０。

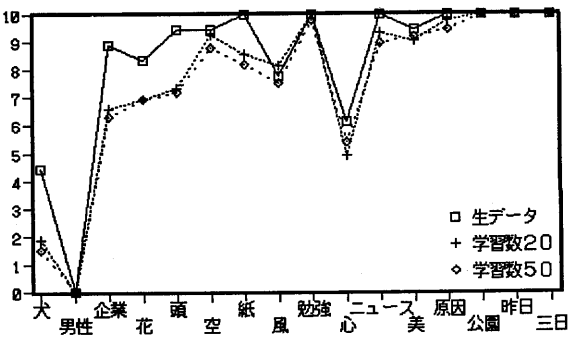

図８　「会う」に対する実測・予測整合度

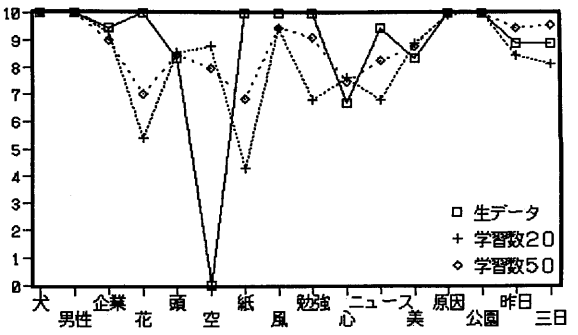

図９　「曇る」に対する実測・予測整合度

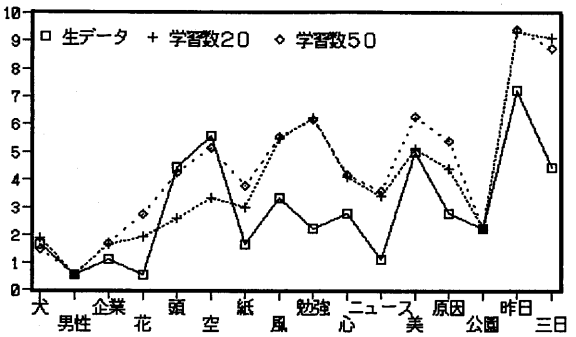

図１０　「残る」に対する実測・予測整合度

各図を見てわかるように、学習数が２０から５０まで増加しても、予測整合度はそれほど変化していない。このなかで、「曇る」に対する主語が「空」の場合、実測整合度は０であるのに、予測整合度は両方とも大きな値を示している。「曇る」のようにほとんどの名詞と共起せず、特定の名詞のみとだけ共起するような場合、この方法を適用することには困難がある。また平均的にこれ以上予測精度を向上させるためには、基準名詞の数を増加することが必要である。いくつの名詞を基準としたら良いかは今後調査したいと思う。

6．まとめ

格助詞「が」を介して名詞が動詞に係る係り受け整合度を内省実験によって求めた。この方法では、テキストから抽出したデータでは出現しにくい係り受けデータが得られ、係り受け全般の性質を把握することができる。しかし、二文節間の係り受け整合度関数を内省によって定める場合、３、で示したような個人による評価の相違をどのように評価するかが問題となる。修飾語や間接目的語の付加によって共起の自然性が変化するという性質は、二文節間の係り受けが、それらだけで決定できないという限界を示している。また、助述成分の変化や格助詞の変更によって生じる共起の自然性の変化も調査する事が必要である。

動詞の主成分分析の結果は、動詞の分類を行う場合、まず、主語となり得る非動物性名詞全般との共起のしやすさ、次に動物の動作であるかどうかで行なうと良いことを示唆している。

整合度テーブルに新たな動詞を追加する場合、その動詞と全ての名詞との整合度を内省によって求めることは、労力がかかる。そこで、その動詞と二つの基準名詞との実測整合度だけから、残りの名詞との整合度を線形予測する試みを行った。その結果、使用する基準名詞は「男性」と「公園」のとき残差が最小となった。また、学習データとして２０個程度の動詞を用いれば、動詞全数学習の場合と同じ程度の予測値が得られた。しかし、「曇る」のように特定の名詞とだけ共起する動詞の整合度を予測するのは困難である。またこれ以上、平均的に予測精度を向上させるには、基準名詞の数を増やさなければならない。いくつぐらい必要かは今後調査したい。

参考文献

（１）尾関：「文節ラチスから最適係り受け構造を選択する多段決定アルゴリズム」信学論、87/12 Vol.J70-D No.12 pp.2621-2629

（２）林：「分類語彙表」国立国語研究所、秀英出版（1966）

（３）大野、浜西：「角川類語新辞典」角川書店（1981）

（４）田中、仁科：「上位／下位関係シソーラスＩＳＡＭＡＰＩの作成［Ｉ］」情処研資ＮＬ64-4 1987 11

（５）仲尾、籾山：「係り関係による単語のクラスタリングの試行」情処研資ＮＬ65-1 1988 3

（６）井ノ上、小倉、森元：「係り受け意味関係の問題点とその考察」電子情報通信学会ＮＬＣ研究会87-25

（７）田中、江原：「日本語「が」格関係に関する考察」情報処理学会第３７回全国大会5B-1
（８）荻野：「名詞と動詞の結合の数理的取り扱い」ソフトウェア文書のための日本語処理の研究－１　３－１０情報処理振興事業協会
（９）「計算機用日本語基本動詞辞書ＩＰＡＬ（Ｂａｓｉｃ　Ｖｅｒｂｓ）解説編」情報処理振興事業協会
（１０）「新用字用語辞典」ＮＨＫ編日本放送出版協会1981
（１１）「新明解国語辞典」三省堂